ILLUMINATING THE DARKNESS

Exploiting Untapped Data and Information Resources in Earth Science

Dr. Rahul Ramachandran

DAAC Manager, GHRC
NASA/MSFC
rahul.ramachandran@nasa.gov
256-961-7620

Global Hydrology Resource Center

Project Team

- Univ. of Alabama in Huntsville
  - Manil Maskey*
  - Xiang Li
- RPI
  - Peter Fox*
  - Stephan Klene
- NASA/GSFC
  - Steve Kempler (Chris Lynnes*)
  - Suhung Shen
  - Chung-Lin Shie

Outline

- Overview of Project
- Use Case Deconstruction
- Initial Results from Data Curation Service

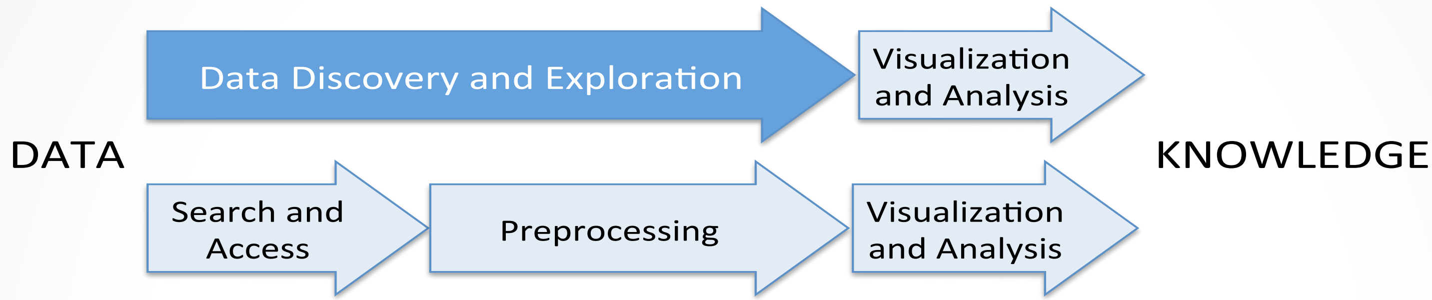
GHRC
Global Hydrology Resource Center

Part 1: Overview

. . .

Global Hydrology Resource Center

Motivation



- Data preparation steps are cumbersome and time consuming
  - Covers discovery, access and preprocessing
- Limitations of current Data and information
  - Searches on data are boolean searches on instrument or geophysical keywords
  - Underlying assumptions that users have sufficient knowledge of the domain vocabulary
  - Lack support for those unfamiliar with the domain vocabulary or the breadth of relevant data available

GHRC Global Hydrology Resource Center

Earth Science Metadata: Dark Resources

- *Dark resources* - information resources that organizations collect, process, and store for regular business or operational activities but fail to utilize for *other* purposes
    - Challenge is to recognize, identify and effectively utilize these dark data stores
- Metadata catalogs contain dark resources consisting of structured information, free form descriptions of data and browse images.
    - EOS Clearing House (ECHO) holds 3666 data collections, 127 million records for individual files and 67 million browse images.

Premise: Metadata catalogs can be utilized *beyond their original design intent* to provide *new data discovery and exploration pathways* to support science and education communities.

GHRC Global Hydrology Resource Center

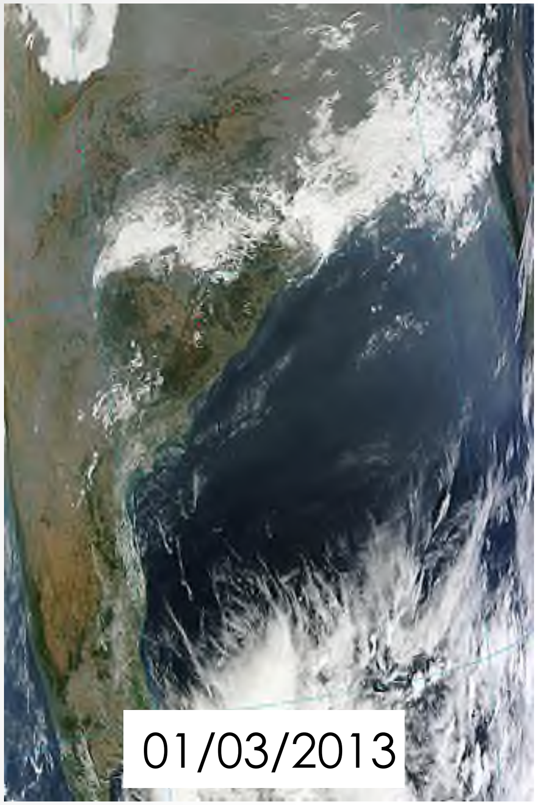
Browse Image Example: Understanding regional air pollution from haze

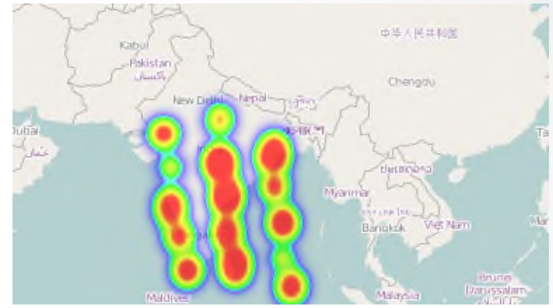
http://rapidfire.sci.gsfc.nasa.gov/cgi-bin/imagery/single.cgi?image=India.A2010345.0510.2km.jpg

- MODIS 2010 image over India which shows modest level haze pollution is used to drive the search
- How often does Haze occur over Indian subcontinent?

Global Hydrology Resource Center

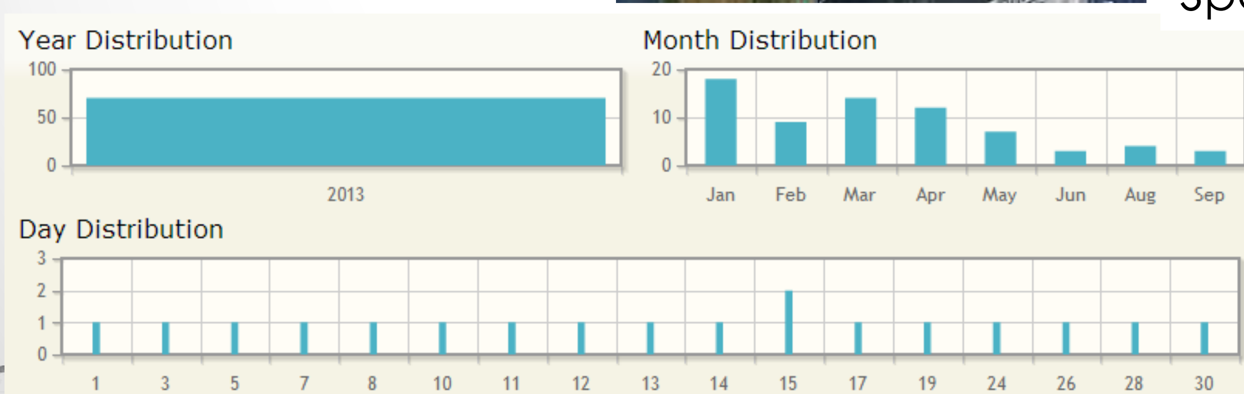
Results: Image Retrieval and Metadata





Spatial Distribution Jan-Sept 13

Spatial Distribution Jan 13



Haze occurs more frequently in Spring than in Summer

Over half a month in January, haze images were observed in the region

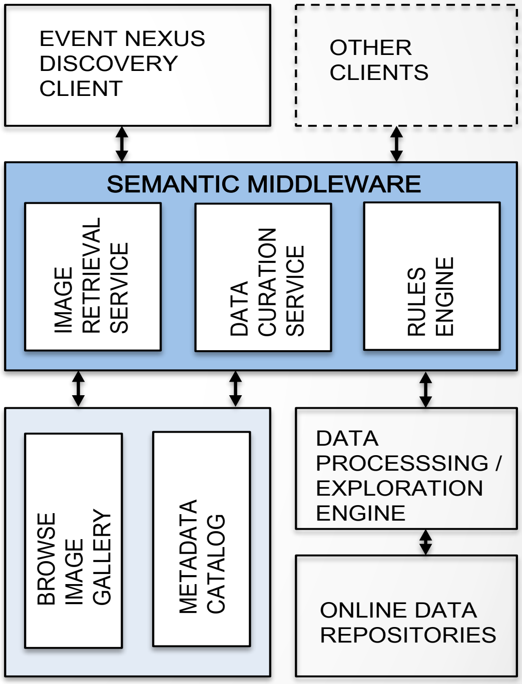
Global Hydrology Resource Center

Goals

- Design a Semantic Middleware Layer (SML) to exploit these metadata resources
  - provide novel data discovery and exploration capabilities that significantly reduce data preparation time.
  - utilize a varied set of semantic web, information retrieval and image mining technologies.
- Design SML as a Service Oriented Architecture (SOA) to allow individual components to be reused and easily integrated into existing NASA's data and information systems.

Specific Objectives

- Three specific semantic middleware core components
  - *Image retrieval service* - uses browse imagery to enable discovery of possible new case studies and granule metadata to present analytics results.
  - *Data curation service* - uses metadata and textual descriptions to find relevant data sets and granules needed to support the analysis of a phenomena or an event.
  - *Semantic rules engine* - automates data preprocessing and exploratory analysis and visualization tasks.
- Demonstrate value using science use cases



Explore pathways to infuse this technology into existing NASA information and data system

GHRC Global Hydrology Resource Center

Science Use Cases

- Dust storms, Volcanic Eruptions, Tropical Storms
- *Volcanic Eruptions*:
  - Emit a variety of gases as well as volcanic ash, which are in turn affected by atmospheric conditions such as winds.
  - Role of Components
    - Image Retrieval Service is used to find volcanic ash events in browse imagery
    - Data Curation Service provides the relevant datasets to support event analysis
    - Rules Engine invokes a Giovanni processing workflow to assemble and compare the wind, aerosol and SO2 data for the vent

GHRC Global Hydrology Resource Center

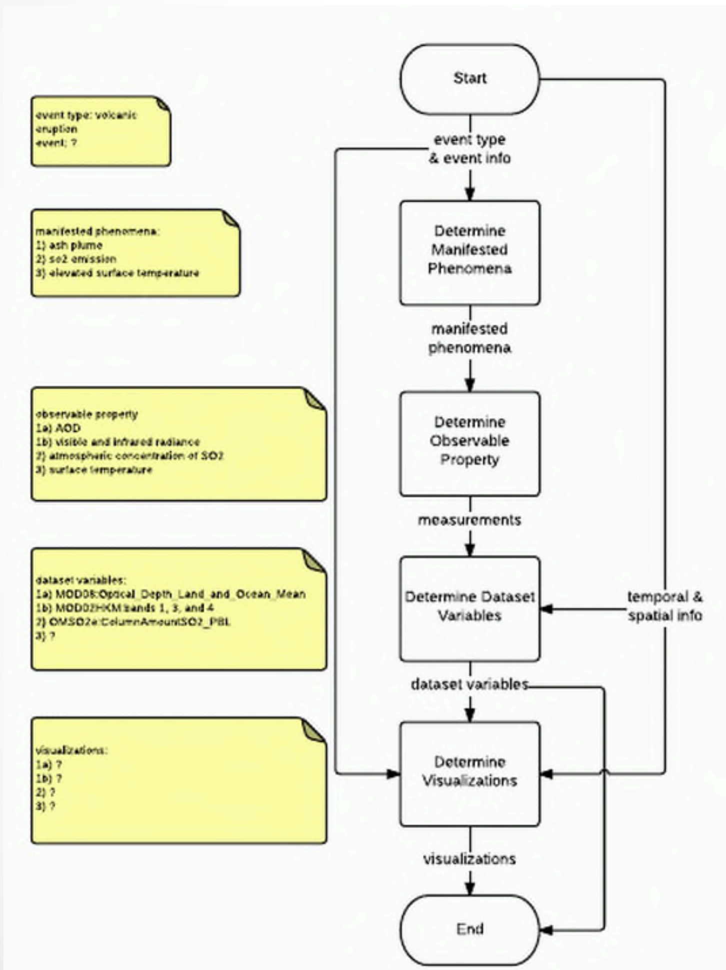
Part 2: Use Case Deconstruction

• • •

Volcanic Eruptions

GHRC
Global Hydrology Resource Center

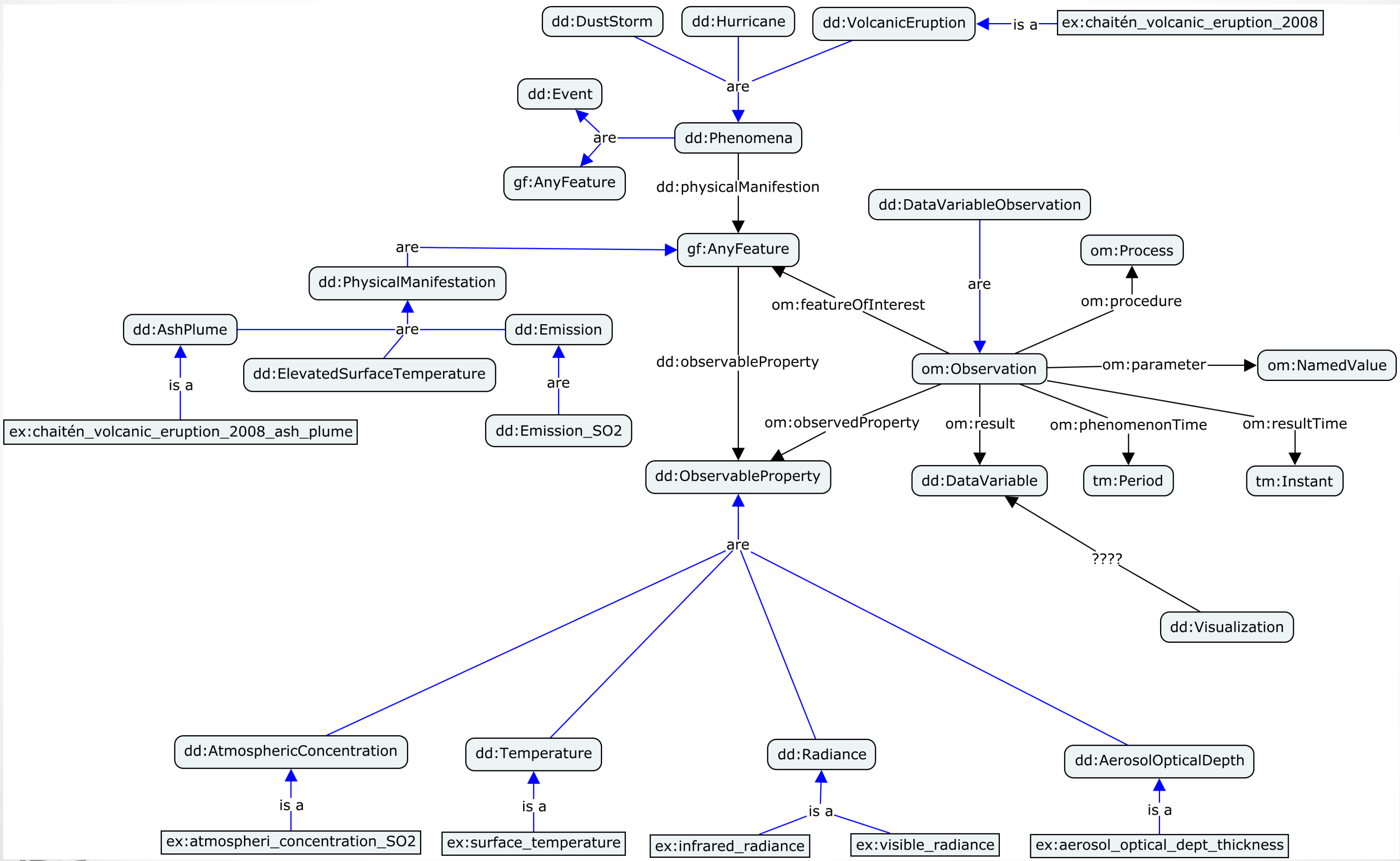
Conceptual Flow and Data Dictionary



Phenomena : As commonly used in weather observing practice, an observable occurrence of particular physical	1. Volcanic eruption 2. Hurricane
Event: Instance of an natural phenomena	1.2008 Chaitén Volcanic eruption, 2.Hurricane Katrina
Physical Manifestation: feature characteristic, the estimation of which is the purpose of an observation	Volcano: Ash plume Hurricane: Wind Fields Eye (Atmospheric Pressure)
Instance (time and space) of physical manifestation	1. 2008 Chaitén ash plume 2. Wind speeds in and around Hurricane Katrina
Measurements (Observable Property): How an instrument observes Phenomena	1. Volcanic Eruption: SO2 Column, Aerosol Optical Depth 2. Hurricane Rainrate Wind speed/direction
Data Set Variable: Representation of the measurement in a data file, variables within an actual data file	OMSO2e:ColumnAmountSO2_PBL MOD08:Optical_Depth_Land_and_Ocean_Mean Precipitation/Visible Frequencies, Pressure

GHRC Global Hydrology Resource Center

Initial Model



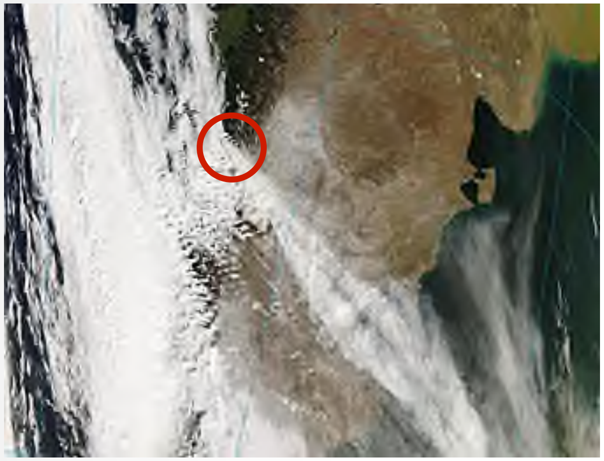
GHRC
Global Hydrology Resource Center

Volcanic Eruption: Chaitén 2008

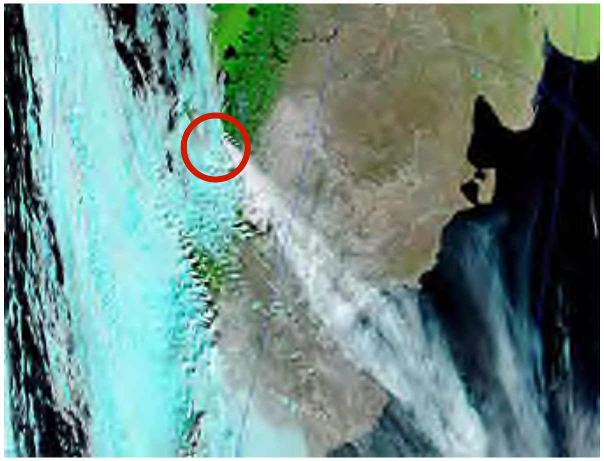
The Chaitén Volcano seen from a commercial flight, October 2008. It was into eruptive phase for the first time in about 9,500 years on the morning of May 2, 2008.

Eruption Time period: May 2 – Nov 2008
Location: Andes region, Chile (-42.832778, -72.645833)

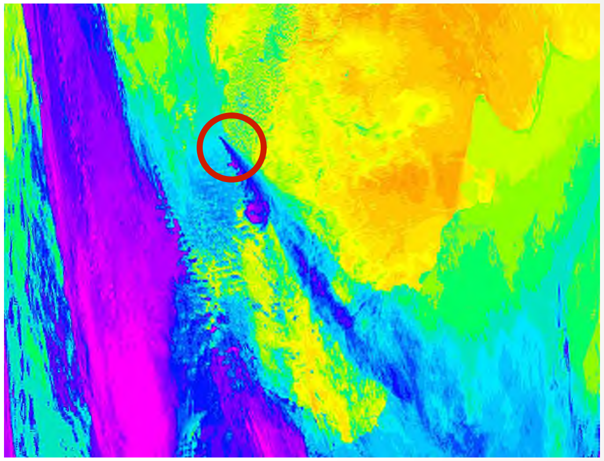
Browse Images



Band 1-4-3 (true color)



Band 7-2-1



LST

Example: MODIS-Aqua 2008-05-03 18:45 UTC

http://lance-modis.eosdis.nasa.gov/cgi-bin/imagery/realtime.cgi?date=2008124

Example Relevant Data

Total SO_2 mass:
e.g. **Chaitén** is 10 (kt) =(kilotons) , (1kt= 1000 metric tons)
ftp://measures.gsfc.nasa.gov/data/s4pa/SO2/MSVOLSO2L4.1/MSVOLSO2L4_v01-00-2014m1002.txt

Daily SO2:
OMI/Aura Sulphur Dioxide (SO2) Total Column Daily L2 Global 0.125 deg
http://disc.sci.gsfc.nasa.gov/datacollection/OMSO2G_V003.html

Calibrated Radiances:
MODIS/Aqua Calibrated Radiances 5-Min L1B Swath 1km
http://dx.doi.org/10.5067/modis/myd021km.006

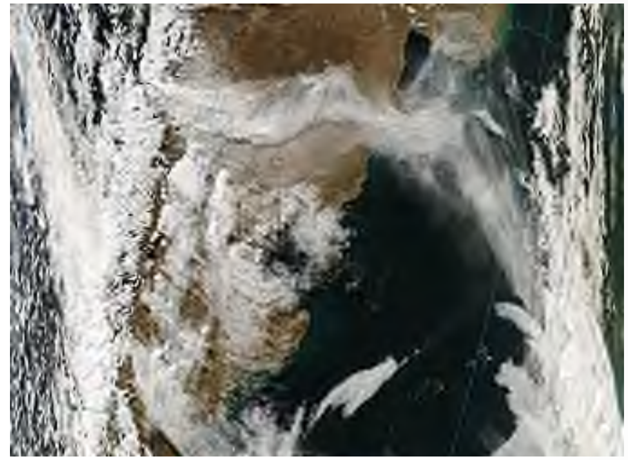
Aerosol Optical Thickness:
MODIS/Aqua Aerosol 5-Min L2 Swath 10km
http://modis-atmos.gsfc.nasa.gov/MOD04_L2/
SeaWiFS Deep Blue Aerosol Optical Depth and Angstrom Exponent Level 2 Data 13.5km
http://disc.gsfc.nasa.gov/datacollection/SWDB_L2_V004.shtml

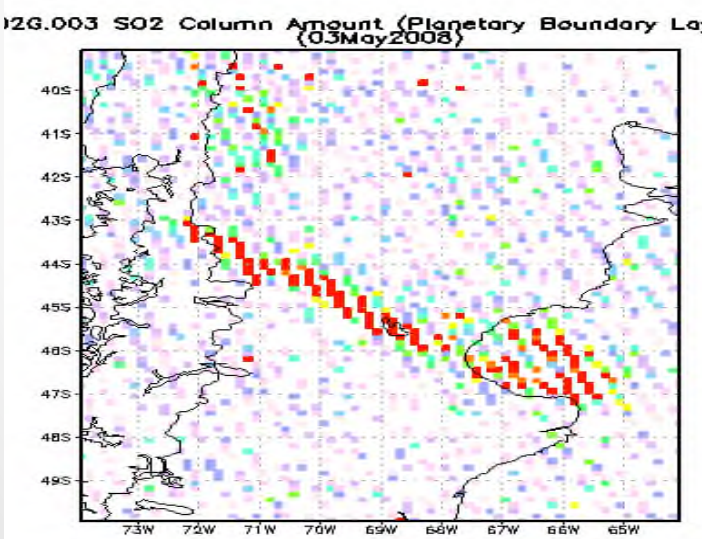
IR Brightness Temperature:
NCEP/CPC 4-km Global (60 deg N - 60 deg S) Merged IR Brightness Temperature Dataset

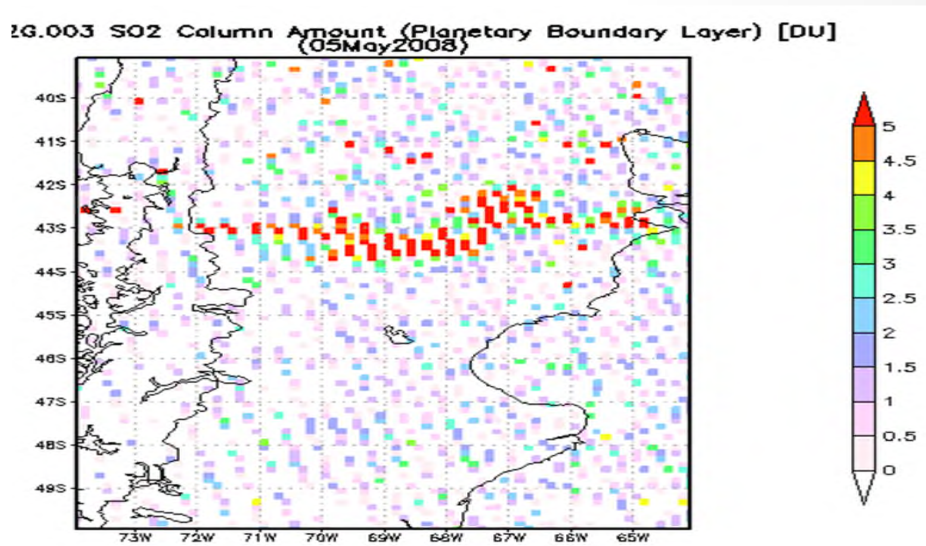
Giovanni SO2 Plots

MODIS-Aqua 2008-05-03 18:45 UTC

MODIS-Aqua 2008-05-05 18:30 UTC



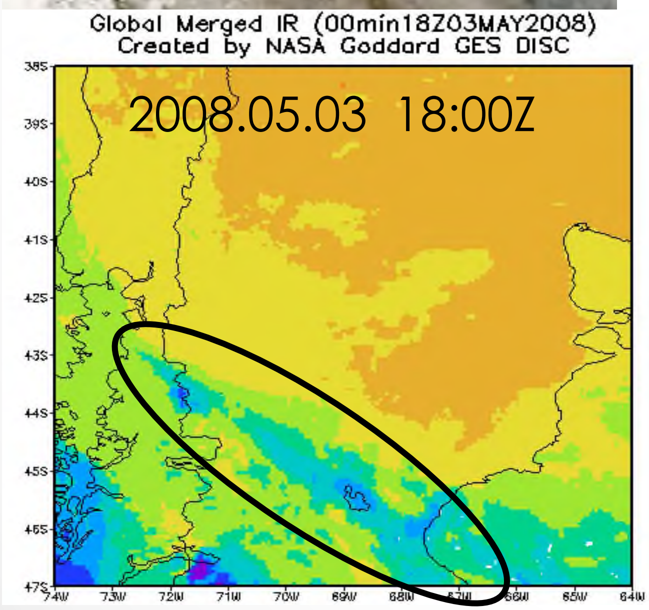




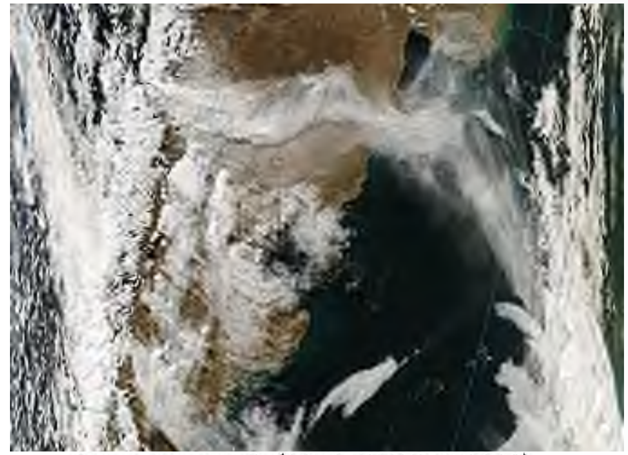
http://gdata2.sci.gsfc.nasa.gov/daac-bin/G3/gui.cgi?instance_id=omil2g

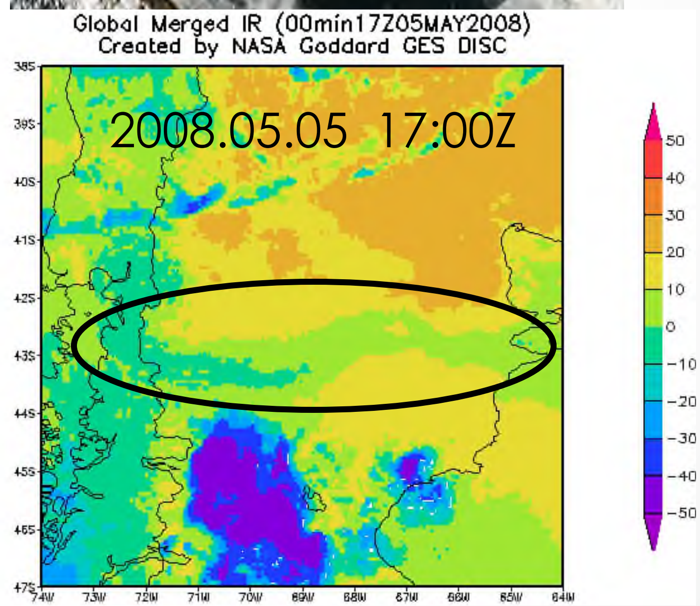
Giovanni Infrared Data Plot

MODIS-Aqua 2008-05-03 18:45 UTC



MODIS-Aqua 2008-05-05 18:30 UTC





http://disc.sci.gsfc.nasa.gov/daac-bin/hurricane_data_analysis_tool.pl

Part C: Data Curation Algorithm for Phenomena

• • •

Initial Results

GHRC
Global Hydrology Resource Center

Data Curation Algorithm Approaches

- Text mining
  - Pros: Don't need to explicitly define the phenomena
  - Cons: Dependent of the truth set; Catalog is dynamic and new data may never get classified
- Ontology Based
  - Pros: Best precision and recall
  - Cons: Labor intensive to build an explicit model
- Information Retrieval
  - Boolean (Faceted) Search
    - Pros: Simple to implement
    - Cons: Phenomena can be complex; User may not know all the right keywords
  - Relevancy Ranking Algorithm
    - Pros: List most relevant data first
    - Cons: Requires a custom algorithm

GHRC Global Hydrology Resource Center

Assumptions/Observations

- Catalog metadata (ECHO) is rich and all metadata records have been tagged with appropriate vocabulary terms (GCMD)
- A phenomena *can be* defined using a bag of keywords using vocabulary terms
  - Information need can be captured by using a broad query
- Keywords (tags) in the metadata and the unstructured text (description) can be used
- Keyword is only used once per metadata record
  - Term frequency does not matter
- Document frequency for keywords can be used
  - Some keywords may occur in many metadata records

Experiment Setup and Approach

- Randomly select 200 sample dataset metadata from ECHO
- Label 200 datasets
  - binary: relevant to phenomena/not relevant to phenomena (Hurricane)
- Compile set of keywords (GCMD) relevant to Hurricane – “bag of words” model
- Filter
  - Spatial filter
  - Temporal resolution
    - “<= daily”
  - 85 datasets filtered out
- Apply algorithms on remaining 115 datasets
  - Jaccard coefficent-based ranking
  - Vector Space Model using Cosine similarity-based ranking

Algorithms

Jaccard Coefficient

$J(A,B) = |A \cap B| / |A \cup B|$

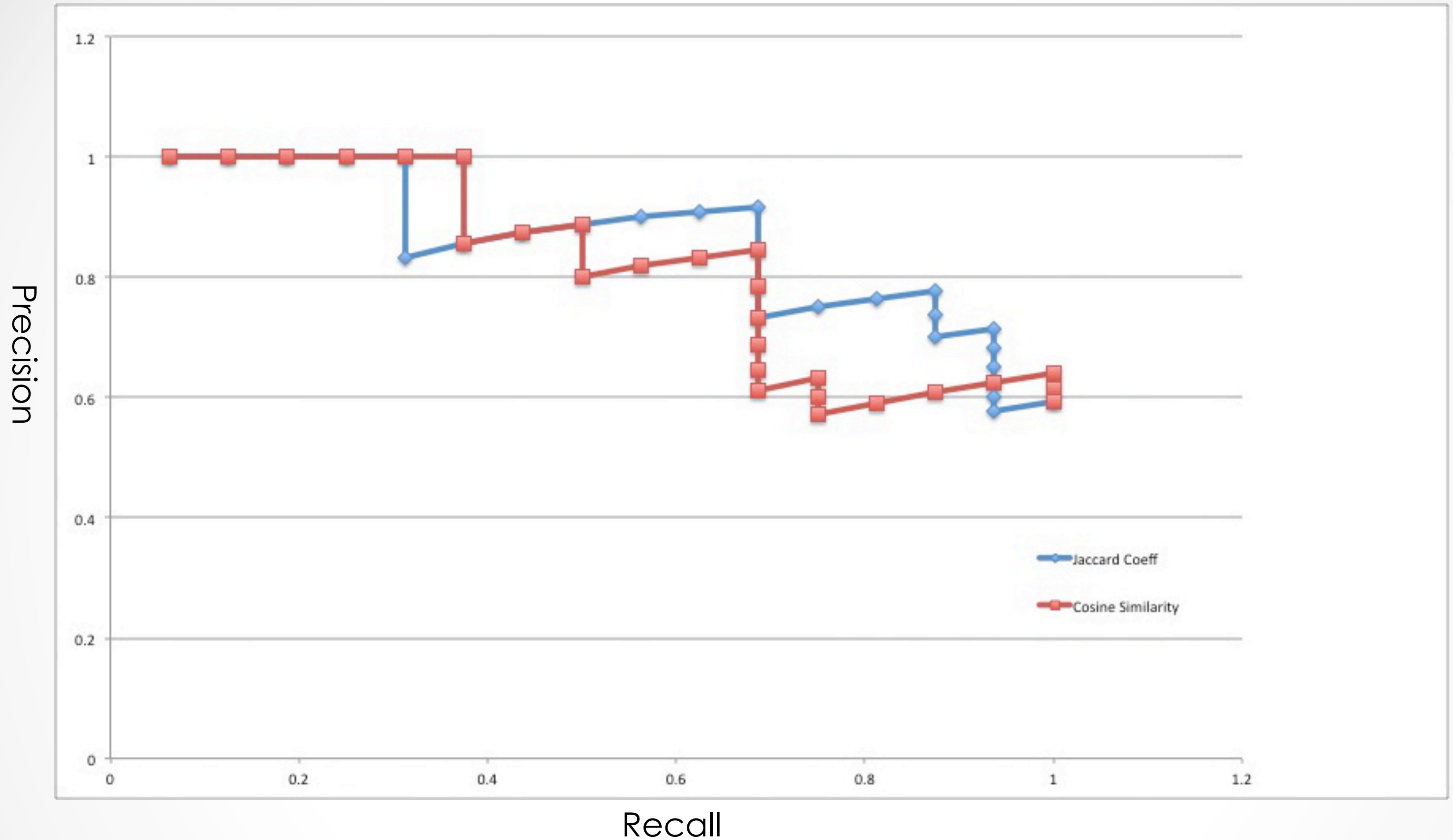
Where:

- A - keywords defining a phenomena
- B - keywords in a given dataset

Vector Space Model

- Determine term frequency (tf): (1 in our case)
- Determine inverse document frequency (idf): number of metadata records that contain the keyword
- Calculate Cosine similarity
  - Sum (tf x idf) for each keyword

Retrieval Results



90 % precision with a 70% recall :
70% of the relevant data are retrieved with 90% precision

Questions

...

Global Hydrology Resource Center